SAND2006-6827P

Temporal Text Analysis of Enron Email using Non-negative PARAFAC

Brett Bader*, *Mike Berry***, and Murray Browne**

*Sandia National Laboratories

**University of Tennessee

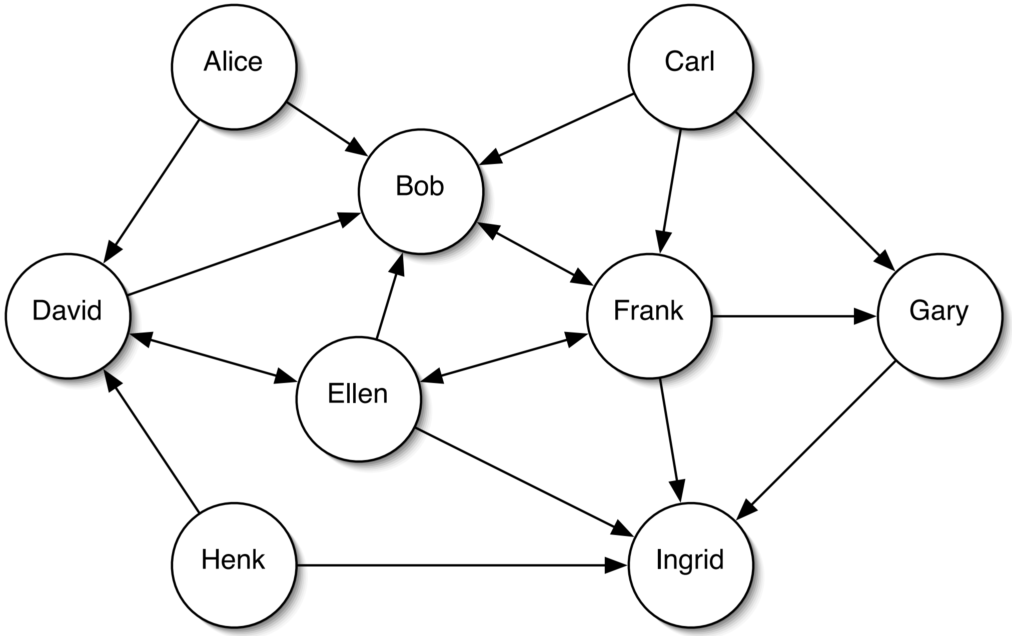
(Slides of joint work for M. Berry to present at a CS Colloquium - UNC)

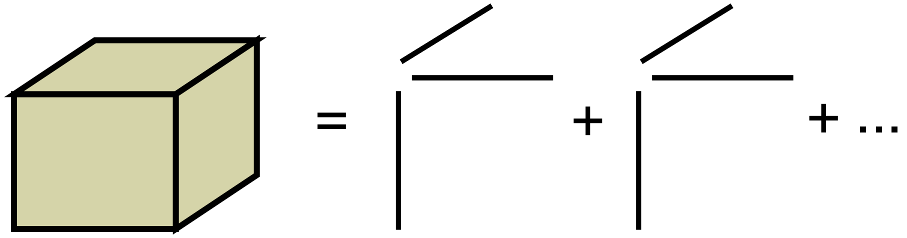
November 1, 2006

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

Using Tensors for Text Analysis

Use PARAFAC to extend LSI to do text analysis

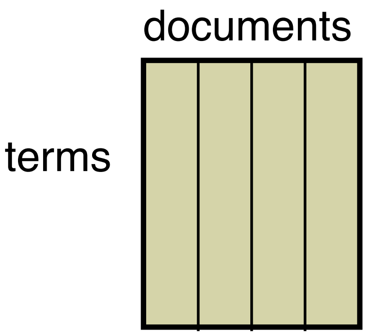




Latent Semantic Indexing

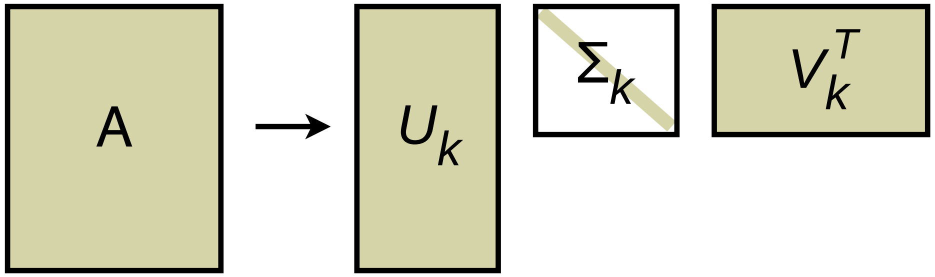
(Deerwester, Dumais, Furnas, Landauer, Harshman, 1990)

Replace term-document matrix with a lower rank matrix that captures “latent” information



Use truncated SVD to compute best rank-*k* matrix

$$A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T \longrightarrow A_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

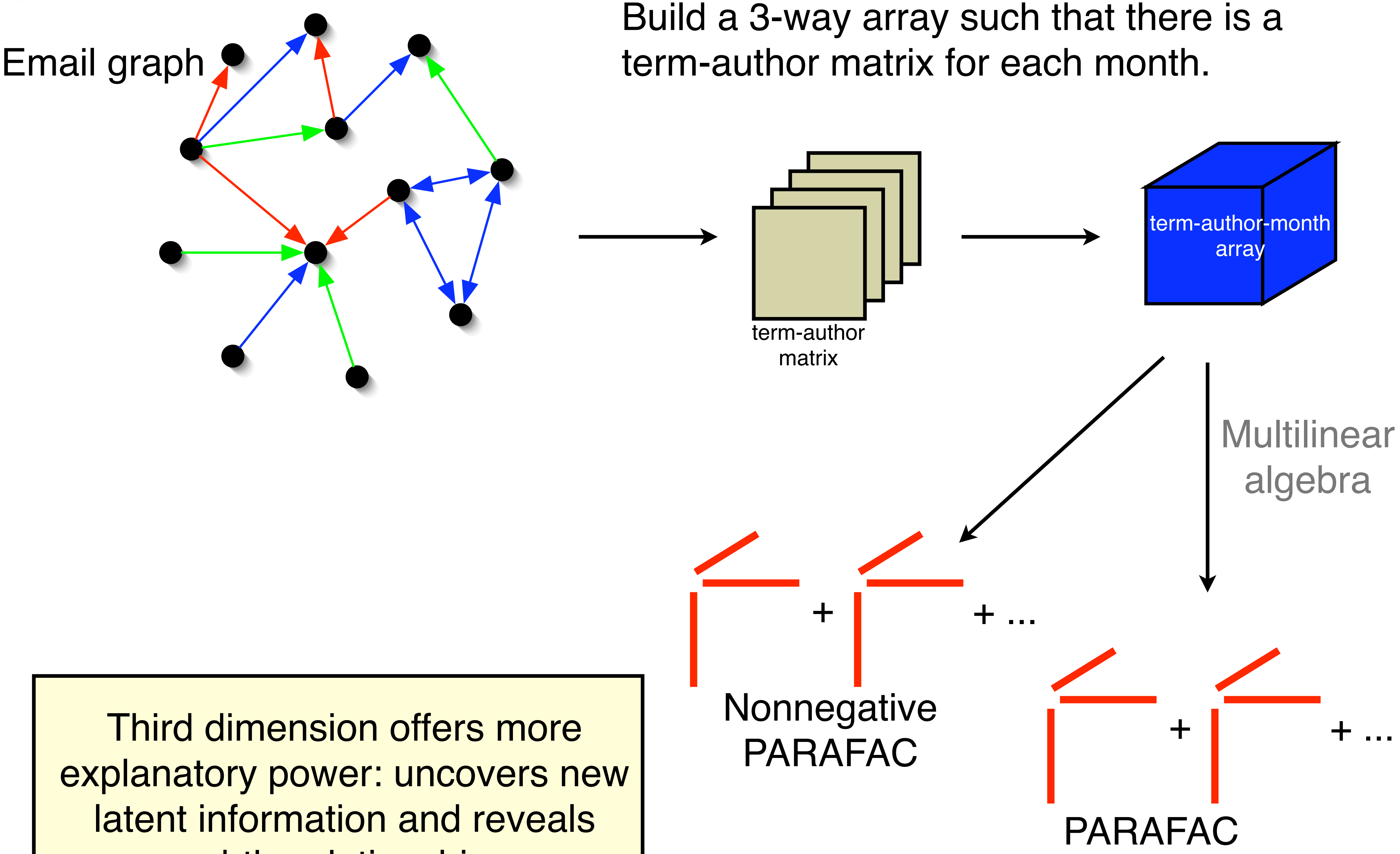


Dimension reduction filters out noise and captures latent information, which improves certain text mining tasks

Example: Search on “car” also finds results on “automobile”

New Paradigm:
Multidimensional Data Analysis

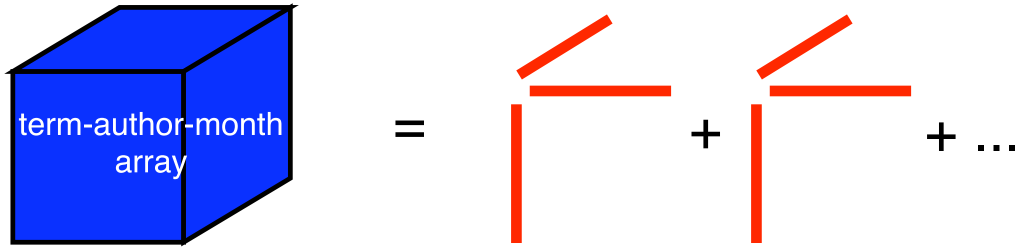
Build a 3-way array such that there is a term-author matrix for each month.

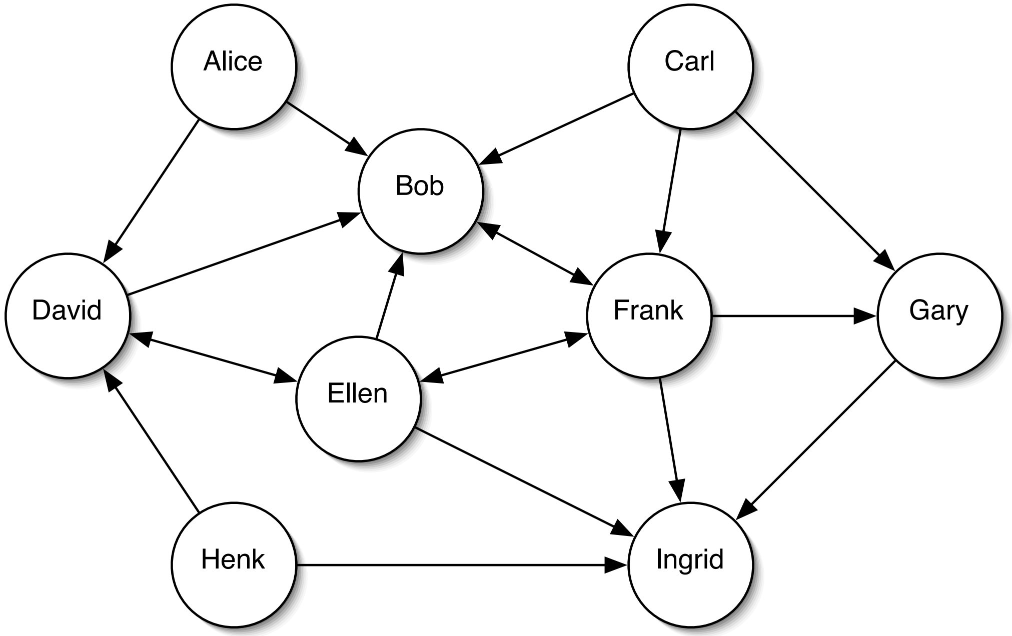


Third dimension offers more explanatory power: uncovers new latent information and reveals subtle relationships

Objective

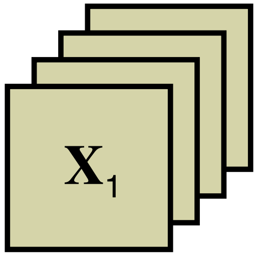
Use PARAFAC to analyze content of email communications over time





Mathematical Notation

- Scalars a
- Vectors $\mathbf{a}$
- Matrices $\mathbf{A}$
- Tensors (3-way array) $\boldsymbol{\mathcal{D}}$ $\boldsymbol{\mathcal{X}}$
    - frontal slices of $\boldsymbol{\mathcal{X}}$: $\mathbf{X}_i$



- Special symbols
    - Kronecker product

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}$$

- Khatri-Rao product (columnwise Kronecker)

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} \mathbf{a}_1 \otimes \mathbf{b}_1 & \dots & \mathbf{a}_n \otimes \mathbf{b}_n \end{bmatrix}$$

- Hadamard product (elementwise)

$$\mathbf{A} * \mathbf{B} = \begin{bmatrix} a_{11}b_{11} & \dots & a_{1n}b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1}b_{m1} & \dots & a_{mn}b_{mn} \end{bmatrix}$$

PARAFAC

- Parallel Factors (Harshman, 1970)
- Also known as CANDECOMP (Carroll & Chang, 1970)

- Many ways to write the model

$$x_{ijk} \approx \sum_{i=1}^{r} a_{ir} b_{jr} c_{kr}$$

$$\boldsymbol{\mathcal{X}} \approx \sum_{i=1}^{r} \mathbf{a}_i \circ \mathbf{b}_i \circ \mathbf{c}_i$$

$$\mathbf{X}^{I \times JK} \approx \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T$$

Matricized array

Typically solved by Alternating Least Squares

Nonnegative PARAFAC

Algorithm adapted from (Morup, 2005) and is based on NMF by (Lee & Seung, 2001)

$$\begin{aligned}\left\| \mathbf{X}^{I\times JK} - \mathbf{A}(\mathbf{C}\odot\mathbf{B})^T \right\|_F &= \left\| \mathbf{X}^{J\times IK} - \mathbf{B}(\mathbf{C}\odot\mathbf{A})^T \right\|_F \\ &= \left\| \mathbf{X}^{K\times IJ} - \mathbf{C}(\mathbf{B}\odot\mathbf{A})^T \right\|_F\end{aligned}$$

Minimize over $\mathbf{A}, \mathbf{B}, \mathbf{C}$ using multiplicative update rule:

$$a_{i\lambda} \leftarrow a_{i\lambda}\frac{(\mathbf{X}^{I\times JK}\mathbf{Z})_{i\lambda}}{(\mathbf{A}\mathbf{Z}^T\mathbf{Z})_{i\lambda}+\epsilon}, \qquad \mathbf{Z} = (\mathbf{C}\odot\mathbf{B})$$

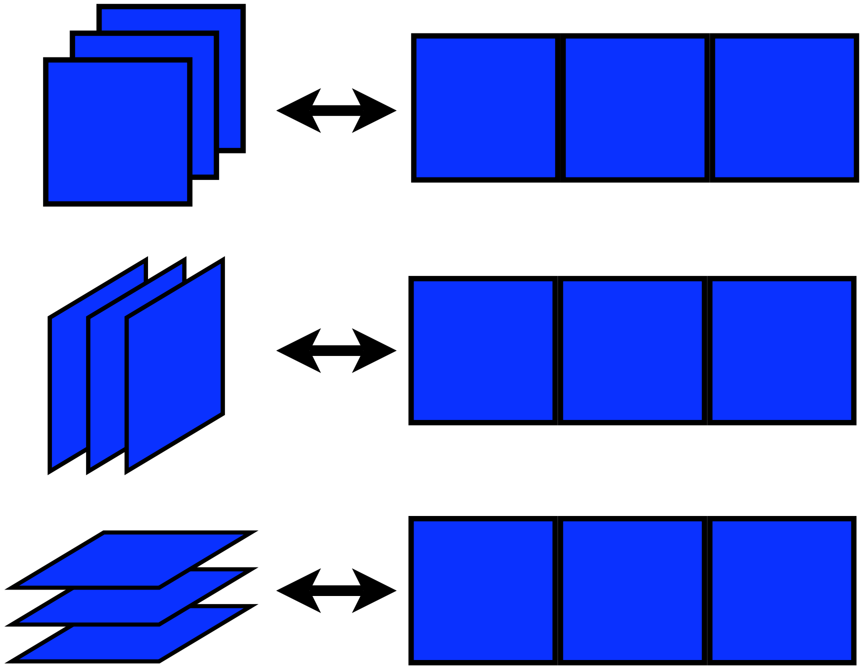
$$b_{j\lambda} \leftarrow b_{j\lambda}\frac{(\mathbf{X}^{J\times IK}\mathbf{Z})_{j\lambda}}{(\mathbf{B}\mathbf{Z}^T\mathbf{Z})_{j\lambda}+\epsilon}, \qquad \mathbf{Z} = (\mathbf{C}\odot\mathbf{A})$$

$$c_{k\lambda} \leftarrow c_{k\lambda}\frac{(\mathbf{X}^{K\times IJ}\mathbf{Z})_{k\lambda}}{(\mathbf{C}\mathbf{Z}^T\mathbf{Z})_{k\lambda}+\epsilon}, \qquad \mathbf{Z} = (\mathbf{B}\odot\mathbf{A})$$

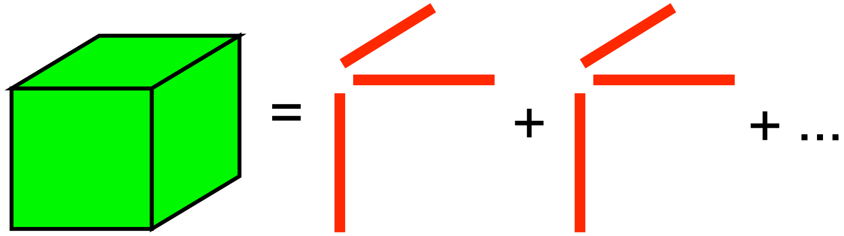
MATLAB Tensor Toolbox

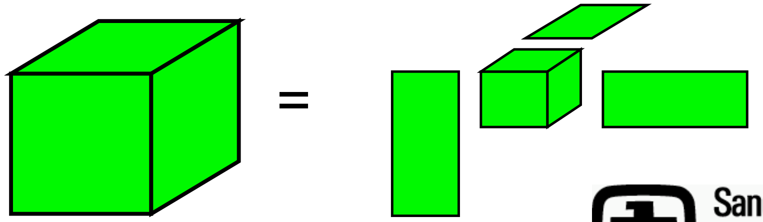
(Bader and Kolda)

- Toolbox extends functionality of Matlab's MDA datatype:
  - Basic operations
  - Convert to/from a matrix
  - Multiplication
    - Tensor
    - Matrix
    - Vector
- Facilitates rapid prototyping of algorithms
  - PARAFAC/CANDECOMP
  - Tucker
  - DEDICOM
- Extensions for a sparse tensor format



$$\mathcal{B} = \mathcal{A} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}$$





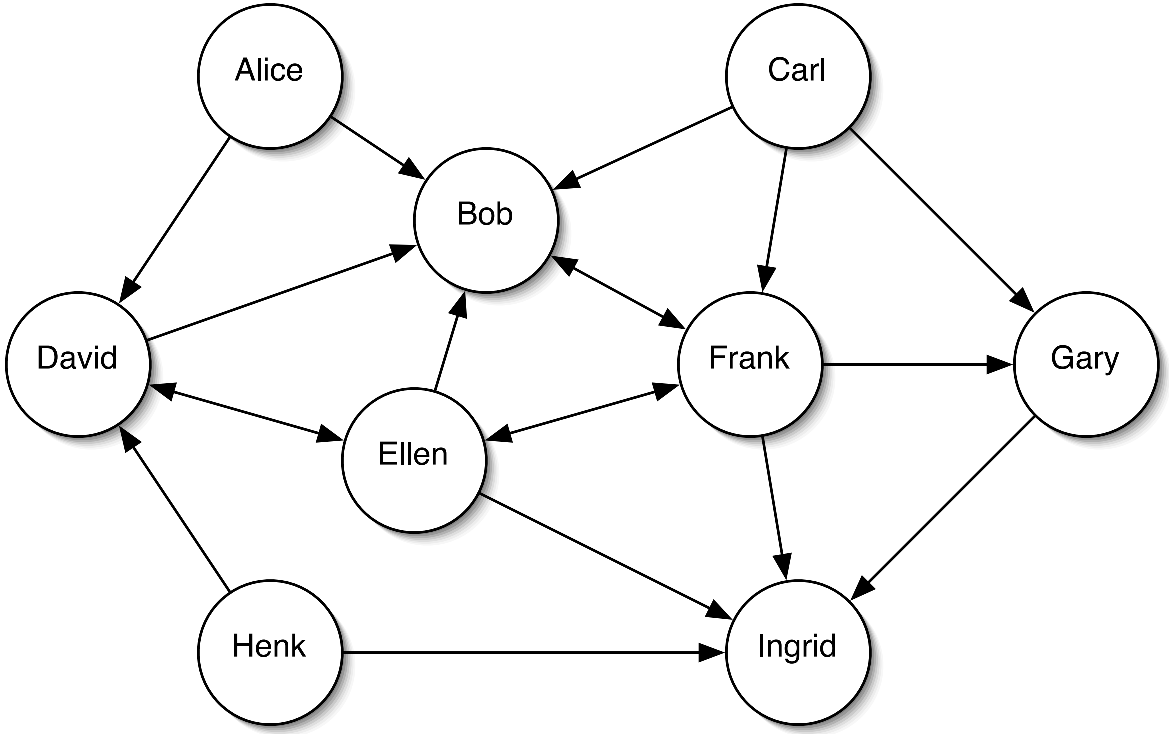
http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox

New Sparse Tensor Class

`sptensor class`

- Coordinate based storage: (i,j,k) indices & values
- Implements many of the same functions as `tensor` class
- Reshape and permute operations handled implicitly with index
- Row / column / slice operations are easy
- Largest problem to date
  - 73,000 x 73,000 x 40,000 sparse tensor
  - 469,000 nonzeros (*very* sparse)
  - Approximate rank-10 PARAFAC model computed in 12 minutes

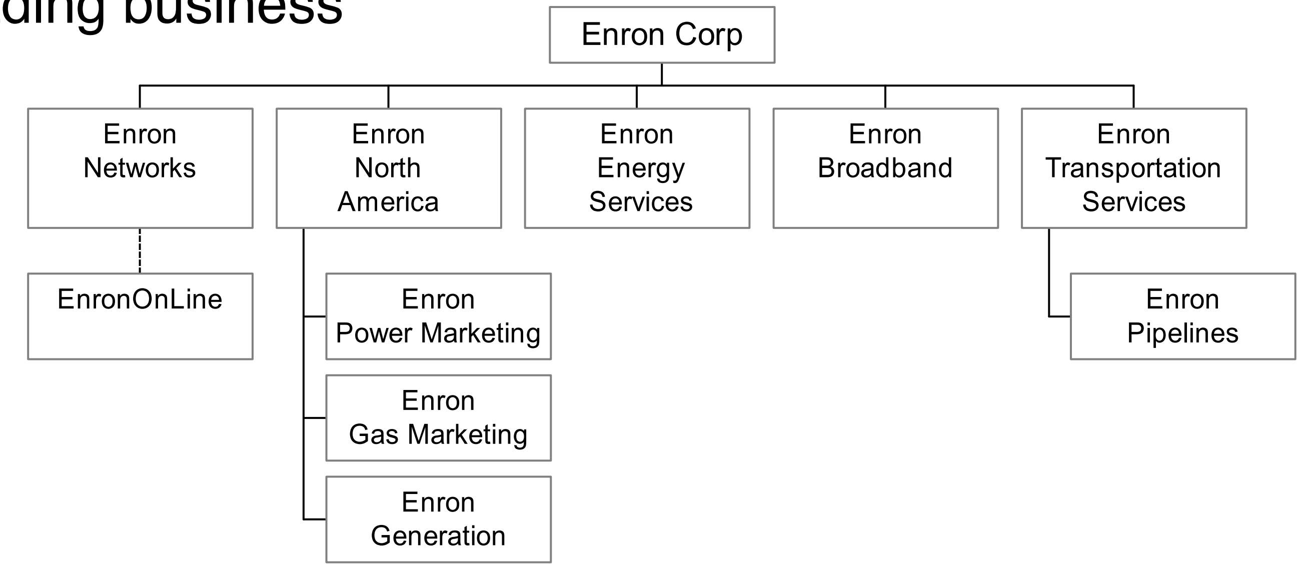
Application: Enron Email Analysis



- Links consist of email communications
- What can we learn about this network strictly from their email communications?
  - Social networks
  - Topics of conversations
  - etc.

Enron Corp.

- U.S. corporation involved with creating energy markets
  - 7th largest by revenue
- EnronOnline: e-trading business
  - natural gas
  - electric power



Enron Corp

Enron Networks

EnronOnLine

Enron North America

Enron Power Marketing

Enron Gas Marketing

Enron Generation

Enron Energy Services

Enron Broadband

Enron Transportation Services

Enron Pipelines

- Investigations
  - U.S. Federal Energy Regulatory Commission (FERC)
    - energy market manipulation
    - involved energy traders
  - U.S. Securities and Exchange Commission (SEC)
    - accounting fraud
    - insider trading

Enron Email Data

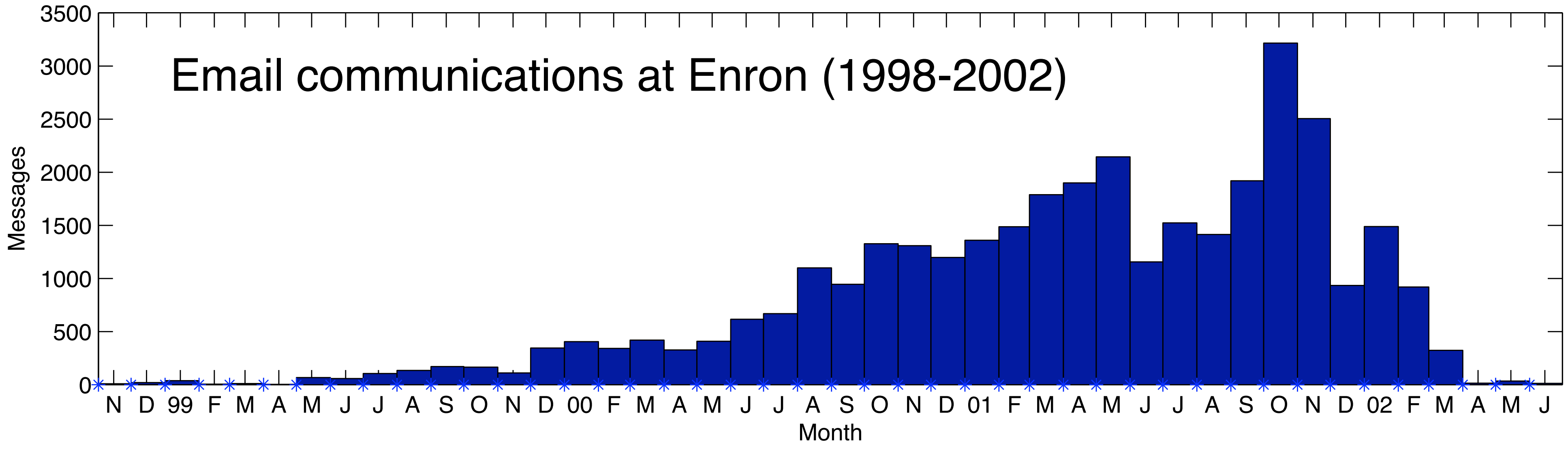
- FERC collected email of ~150 employees as evidence
  - Included emails saved in inbox, sent items, deleted items, and all other folders

- Released to the public in 2002 by FERC as part of their investigation
  - To/from, date, subject, body
  - Attachments and some names/emails removed
  - Approx. 500,000 email messages

- Research uses:
  - Email classification
  - Natural language processing
  - Organizational theory/behavior
  - Social network analysis

Smaller Enron Data Set

We used a smaller data set prepared by Priebe et al.
34,427 emails among 184 employees in 2001
and added 13 “interesting” employees



- Email folders collected at one point in time
- Shape of histogram depends on:
  - How far back employees kept emails
  - Employment history of individual
- Limited information on former employees
  - No org chart
- Collected nouns from each author at each month

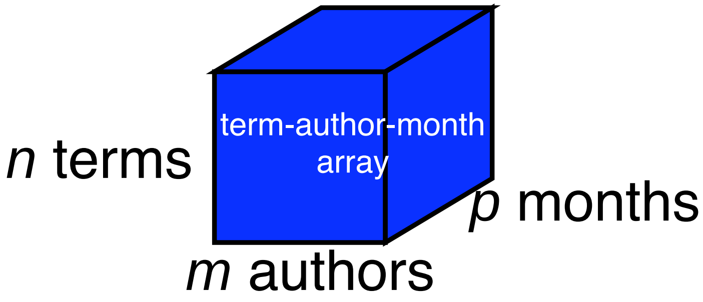
Text Analysis Experiment

- Term-author-time array
  - Sparse tensor of size 69157 x 197 x 12 (1,042,202 nonzeros)

- Term weighted adjacency array

- Models:
  - PARAFAC
  - Non-negative PARAFAC

Term Weighting



Weighted frequency: $x_{ijk} = l_{ijk} g_i a_j$

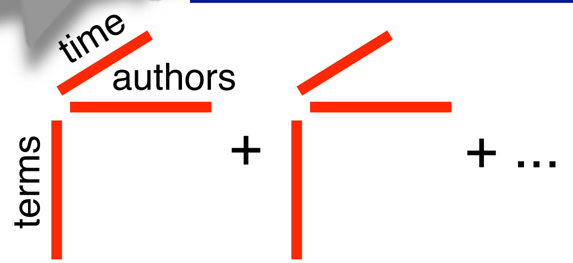
Log local weight: $l_{ijk} = \log(1 + f_{ijk})$

Entropy global weight: $g_i = 1 + \sum_{j=1}^{n} \frac{h_{ij} \log h_{ij}}{\log n}$

where $h_{ij} = \frac{\sum_k f_{ijk}}{\sum_{jk} f_{ijk}}$

Author normalization: $a_j = \frac{1}{\sqrt{\sum_{i,k} (l_{ijk} g_i)}}$

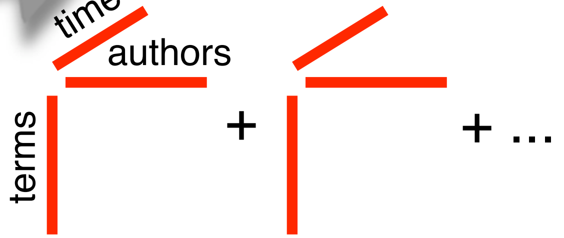
Conversation Topics of Employees



California Energy

Nouns		Employees	
Score	Noun	Score	Employee
Factor 4			
0.094	california	0.497	James Steffes (james.steffes) VP Government Affairs
0.087	dasovich	0.430	Steven Kean (steven.kean) VP Chief of Staff
0.079	jeff	0.413	Jeff Dasovich (jeff.dasovich) Employee Government Relationship Executive
0.077	shapiro	0.319	Richard Sanders (richard.sanders) VP Enron Wholesale Services
0.076	steffes	0.219	Richard Shapiro (richard.shapiro) VP Regulatory Affairs
0.075	richard	0.194	Elizabeth Sager (elizabeth.sager) VP and Asst Legal Counsel ENA Legal
0.073	kean	0.187	Mark Haedicke (mark.haedicke) Managing Director ENA Legal
0.072	edison	0.171	Drew Fossum (drew.fossum) VP Transwestern Pipeline Company (ETS)?
0.067	utilities	0.152	Philip Allen (phillip.allen) VP West Desk Gas Trading
0.066	power	0.134	Kay Mann (kay.mann) Lawyer
0.065	sanders	0.125	Mark Taylor (mark.taylor) Manager Financial Trading Group ENA Legal
0.064	mara	0.100	John Arnold (john.arnold) VP Financial Enron Online
0.063	james	0.097	Margaret Carson (margaret.carson) Employee Corporate and Environmental Policy*
0.062	development	0.095	Kevin Presto (kevin.presto) VP East Power Trading
0.061	governor	0.085	Vince Kaminski (vince.kaminski) Manager Risk Management Head
0.061	vicki	0.081	David Delainey (david.delainey) CEO ENA and Enron Energy Services
0.058	energy	0.072	Rick Buy (rick.buy) Manager Chief Risk Management Officer
0.057	kaufman	0.069	Sara Shackleton (sara.shackleton) Employee ENA Legal
0.055	mccubbin	0.060	Kate Symes (kate.symes) Employee
0.055	kingerski	0.059	Gerald Nemec (gerald.nemec) N/A
0.055	utility	0.055	Larry Campbell (larry.campbell) Employee Senior Specialist
0.055	sharp	0.055	Michael Grigsby (mike.grigsby) Director West Desk Gas Trading
0.055	market	0.054	Dan Hyvl (dan.hyvl) Employee
0.054	electricity	0.054	Mike McConnell (mike.mcconnell) Executive VP* Global Markets
0.054	alan	0.050	Bruce Lundstrom (bruce.lundstrom) N/A

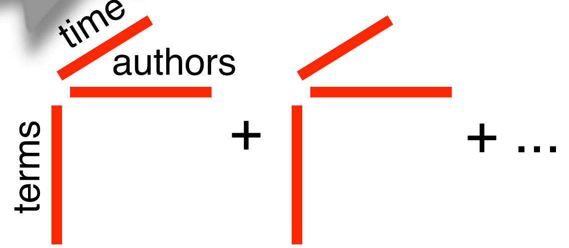
Conversation Topics of Employees



Enron crisis and downfall

Nouns		Employees	
Score	Noun	Score	Employee
Factor 12			
0.072	amat	0.790	Teb Lokey (teb.lokey) Manager Regulatory Affairs
0.063	skilling	0.480	Shelley Corman (shelley.corman) VP Regulatory Affairs
0.062	billion	0.256	Darrell Schoolcraft (darrell.schoolcraft) Employee Gas Control (ETS)
0.055	hng	0.176	Tana Jones (tana.jones) Employee Financial Trading Group ENA Legal
0.054	jeff	0.165	Louise Kitchen (louise.kitchen) President Enron Online
0.051	profits	0.076	Daron Giron (c..giron) Employee
0.051	percent	0.073	Phillip Love (m..love) N/A
0.050	california	0.047	Mark Whitt (mark.whitt) Director Marketing
0.050	electricity	0.031	Michael Grigsby (mike.grigsby) Director West Desk Gas Trading
0.049	blair	0.027	James Derrick (james.derrick) In House Lawyer
0.049	gest	0.026	Jay Reitmeyer (jay.reitmeyer) Associate Eastern Rockies Natural Gas Trader
0.047	teb	0.024	Lynn Blair (lynn.blair) Employee Northern Natural Gas Pipeline (ETS)
0.046	wall	0.024	Benjamin Rogers (benjamin.rogers) Employee Associate
0.043	lokey	0.021	Bruce Lundstrom (bruce.lundstrom) N/A
0.043	energy	0.021	Steven Kean (j..kean) VP Chief of Staff
0.043	customers	0.020	Stacey White (w..white) N/A
0.042	power	0.020	Jeff Dasovich (jeff.dasovich) Employee Government Relationship Executive
0.042	lay	0.019	James Steffes (d..steffes) VP Government Affairs
0.042	deregulation	0.018	John Arnold (john.arnold) VP Financial Enron Online
0.041	virgilio	0.018	Joe Quenet (joe.quenet) Trader
0.041	coale	0.018	Jeffrey Shankman (a..shankman) President Enron Global Markets
0.041	street	0.017	xxx Harris (j.harris) xxx
0.041	plants	0.013	Kim Ward (kim.ward) Manager West Gas Origination
0.041	million	0.011	Kenneth Lay (kenneth.lay) CEO
0.041	stock	0.010	Bill Williams (bill.williams) xxx

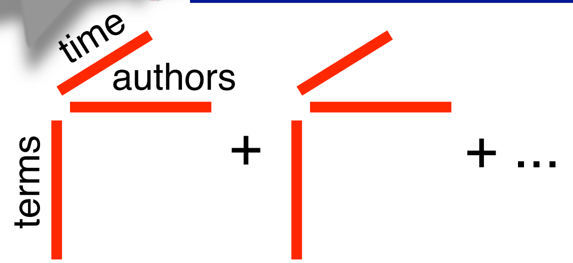
Conversation Topics of Employees



Education and hiring student interns

Nouns		Employees	
Score	Noun	Score	Employee
Factor 17			
0.352	kaminski	0.996	Vince Kaminski (j..kaminski) Manager Risk Management Head
0.248	krishna	0.052	Stanley Horton (stanley.horton) President Enron Gas Pipeline
0.238	pinnamaneni	0.035	Jeffery Skilling (jeff.skilling) CEO
0.237	vince	0.033	Vince Kaminski (j.kaminski) Manager Risk Management Head
0.233	krishnarao	0.025	Vince Kaminski (vince.kaminski) Manager Risk Management Head
0.213	marcusevansch	0.024	Joannie Williamson (joannie.williamson) Executive Assistant
0.171	visa	0.019	Kimberly Watson (kimberly.watson) Employee Transwestern Pipeline Company (ETS)
0.156	instructor	0.017	Thomas Martin (a..martin) VP
0.149	clare	0.016	Louise Kitchen (louise.kitchen) President Enron Online
0.138	pallavi	0.012	Sandeep Kohli (sandeep.kohli) N/A
0.128	humphreys	0.011	Jeff King (jeff.king) Manager
0.125	pregnancy	0.011	Cooper Richey (cooper.richey) Manager
0.122	instructors	0.011	Bruce Lundstrom (bruce.lundstrom) N/A
0.121	vkamins	0.010	Philip Allen (phillip.allen) VP West Desk Gas Trading
0.117	agenda		
0.117	overstayed		
0.117	courses		
0.113	intervene		
0.109	fitzgerald		
0.102	confrontational		
0.101	behalf		
0.099	course		
0.095	humanitarian		
0.095	complicated		
0.094	almighty		

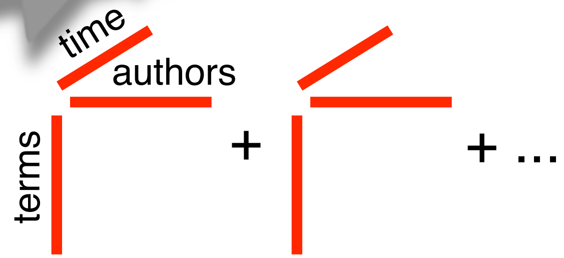
Conversation Topics of Employees



College Football

Nouns		Employees	
Score	Noun	Score	Employee
Factor 21			
0.189	bcs	0.737	Matthew Motley (matt.motley) Director
0.155	byu	0.606	Randall Gay (l..gay) West Desk Gas Trading
0.120	sooners	0.143	Craig Dean (craig.dean) Trader
0.119	frommelt	0.119	Mark Taylor (e.taylor) Manager Financial Trading Group ENA Legal
0.117	nebraska	0.091	Clint Dean (clint.dean) xxx
0.109	bowl	0.057	Kam Keiser (kam.keiser) Employee Gas
0.104	pooky	0.054	Eric Bass (eric.bass) Trader Texas Desk Gas Trading
0.102	gay	0.049	Thomas Martin (a..martin) VP
0.099	oklahoma	0.048	Cooper Richey (cooper.richey) Manager
0.097	big	0.045	Don Baughman (don.baughman) Trader
0.095	cougars	0.044	John Griffith (john.griffith) xxx
0.091	kathleen	0.044	Daren Farmer (j..farmer) Manager Logistics Manager
0.090	horns	0.044	Jim Schwieger (jim.schwieger) Trader Texas Desk Gas Trading
0.088	rooting	0.042	Kevin Hyatt (kevin.hyatt) Director Asset Development TW Pipeline Business (ETS)
0.086	fiesta	0.042	Albert Meyers (albert.meyers) Employee Specialist
0.085	tennessee	0.040	Bill Rapp (bill.rapp) N/A
0.085	texas	0.040	Michael Maggi (mike.maggi) Director
0.083	grigsby	0.039	Stanley Horton (stanley.horton) President Enron Gas Pipeline
0.081	longhorn	0.036	Cara Semperger (cara.semperger) Employee Senior Analyst Cash
0.081	oregon	0.033	Jeff King (jeff.king) Manager
0.080	longhorns	0.032	Sandra Brawner (f..brawner) Director
0.077	espn	0.031	Tom Donohoe (tom.donohoe) Trader Central Desk Gas Trading
0.077	miami	0.029	Jay Reitmeyer (jay.reitmeyer) Associate Eastern Rockies Natural Gas Trader
0.077	stanford	0.027	Jane Tholt (m..tholt) VP West Desk Gas Trading
0.077	large	0.027	Matthew Lenhart (matthew.lenhart) Analyst West Desk Gas Trading

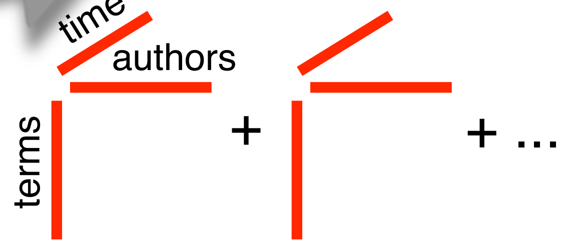
Conversation Topics of Employees

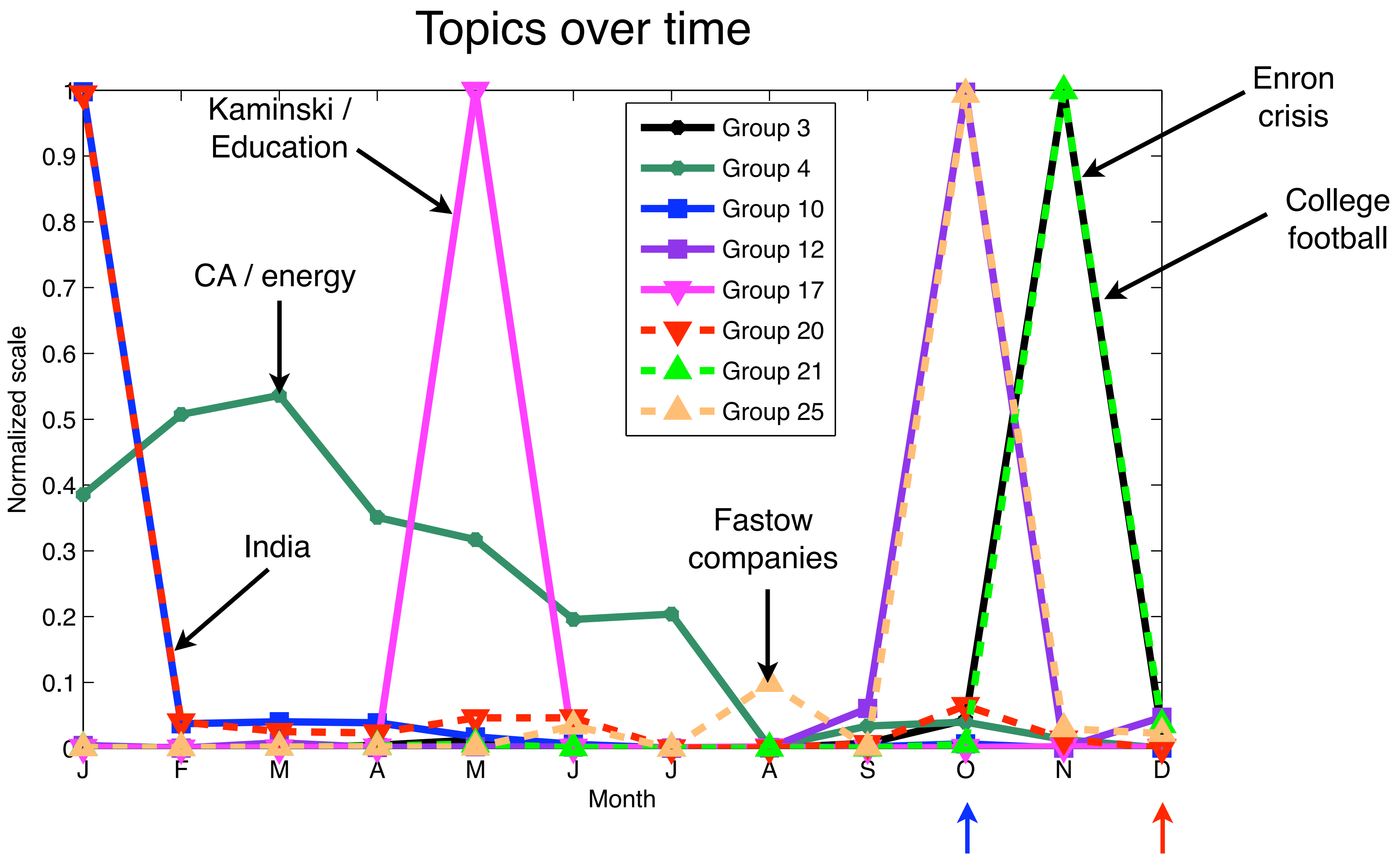


Andrew Fastow's companies and
CA project names leading to blackouts

Nouns		Employees	
SCORE	NOUN	SCORE	EMPLOYEE
Factor 25			
0.136	essie	0.811	Mary Fischer (mary.fischer) Employee
0.126	jeff	0.440	xxx Harris (j.harris) xxx
0.100	locklear	0.269	John Hodge (t..hodge) Managing Director
0.090	caminos	0.128	Stacey White (w..white) N/A
0.089	fischer	0.098	James Derrick (james.derrick) In House Lawyer
0.088	harrier	0.085	Steven Kean (j..kean) VP Chief of Staff
0.087	herrold	0.083	Jay Reitmeyer (jay.reitmeyer) Associate Eastern Rockies Natural Gas Trader
0.083	million	0.076	Louise Kitchen (louise.kitchen) President Enron Online
0.083	october	0.066	Jeffrey Shankman (a..shankman) President Enron Global Markets
0.083	pronghorn	0.055	Clay Harris (clay.harris) N/A
0.082	raptor	0.052	Benjamin Rogers (benjamin.rogers) Employee Associate
0.082	mary	0.050	Harpreet Arora (harry.arora) VP
0.082	kate	0.046	Bruce Lundstrom (bruce.lundstrom) N/A
0.081	facundo	0.044	Richard Sanders (b..sanders) VP Enron Wholesale Services
0.080	eloise	0.036	John Lavorato (john.lavorato) CEO Enron America
0.079	documents	0.034	Peter Keavey (f..keavey) Employee
0.078	vicsandra	0.032	Mark Haedicke (e..haedicke) Managing Director ENA Legal
0.077	roadrunner	0.032	Rick Buy (rick.buy) Manager Chief Risk Management Officer
0.073	walker	0.031	Marie Heard (marie.heard) Senior Legal Specialist ENA Legal
0.072	grizzly	0.031	Phillip Love (m..love) N/A
0.072	booklet	0.030	Sally Beck (sally.beck) COO
0.071	lynda	0.023	John Zufferli (john.zufferli) VP Canada Gas Trading
0.068	entity	0.022	Sara Shackleton (sara.shackleton) Employee ENA Legal
0.067	leesa	0.021	Debra Perlingiere (debra.perlingiere) Legal Specialist ENA Legal
0.067	porcupine	0.020	John Hodge (john.hodge) Managing Director East Gas/Origination*

Temporal Patterns





Enron crisis breaks; investigation begins

Filed for bankruptcy

Summary

- Nonnegative Tensor Factorization

- Novel approach to email surveillance using PARAFAC
  - Topics of conversation
  - Employees involved
  - Communication patterns over time

More Information

bwbader@sandia.gov
http://www.cs.sandia.gov/~bwbader/

- MATLAB Tensor Toolbox version 2.0:
  - http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox
  - Tech report SAND2004-5189 available on website
  - Paper to appear in ACM Trans. Math. Softw.